Classification and Exploration of 3D Protein Domain Interactions Using Kbdock

Anisah W Ghoorah, Marie-Dominique Devignes, Malika Smaïl-Tabbone, David Ritchie

Classification and Exploration of 3D Protein Domain Interactions using Kbdock

Anisah W. Ghoorah,

Marie-Dominique Devignes ,

Malika Smaïl-Tabbone ,

David W. Ritchie*

Department of Computer Science and Engineering, University of Mauritius, 80837 Reduit, Mauritius
CNRS, LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France
Inria Nancy – Grand Est, 54600 Villers-lès-Nancy, France
University of Lorraine, LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France

dave.ritchie@inria.fr, +33 3 83 59 30 45

Keywords

structural biology;
structural homology;
protein domains;
protein domain family;
domain-domain interactions;
domain-peptide interactions;
domain family interactions;
domain family binding sites.

Running Head

Kbdock 3D Protein Domain Interactions

Abstract

Comparing and classifying protein domain interactions according to their three-dimensional (3D) structures can help to understand protein structure-function and evolutionary relationships. Additionally, structural knowledge of existing domain-domain interactions can provide a useful way to find structural templates with which to model the 3D structures of unsolved protein complexes. Here we present a straight-forward guide to using the "Kbdock" protein domain structure database and its associated web site for exploring and comparing protein domain-domain interactions (DDIs) and domain-peptide interactions (DPIs) at the Pfam domain family level. We also briefly explain how the Kbdock web site works, and we provide some notes and suggestions which should help to avoid some common pitfalls when working with 3D protein domain structures.

1 Introduction

Protein-protein interactions (PPIs) are fundamental biophysical interactions. Consequently, comparing and classifying PPIs at the molecular level can enrich our understanding of many biological processes. In order to relate the structure and function of different proteins in a systematic way, PPIs are often described in terms of domain-domain interactions (DDIs) because protein domains may often be identified as structural and functional units. While many PPIs may involve rapid or transitory interactions *in vivo*, many others involve the formation of long-lasting three-dimensional (3D) protein-protein complexes. Under favourable conditions, these 3D structures may be observed at low resolution using cryo-electron microscopy, or they may be captured at atomic resolution using X-ray crystallography or nuclear magnetic resonance spectroscopy. These complexes may consist of homo-dimers or higher order homo-multimers, or they may involve heteromeric interactions between different protein chains. While homo-interactions are observed relatively often in crystal structures, most processes of biological interest involve hetero interactions, and the corresponding structures are normally much more difficult to determine experimentally and to predict computationally [1]. Consequently, although the number of solved 3D protein structures appears to be growing exponentially [2], there is an equally growing need to be able to classify and analyse the structural repertoire of known hetero PPIs using computational modeling and analysis techniques.

Three widely used domain definitions are Pfam [3], SCOP [4], and CATH [5]. Pfam defines domains using multiple sequence alignments in order to identify families of sequences which will often correspond to distinct functional and structural regions. The SCOP and CATH classifications use both sequence and structural similarities to collect protein domains in a hierarchical system of related domain families. However, these two classifications are constructed using different sequence-based and structure-based alignment tools, and they both require the use of considerable human expertise to deal with novel structures which cannot be classified automatically. We therefore choose to work directly with the sequence-based Pfam classification which does not attempt to define a complex structural hierarchy like SCOP and CATH, but which nonetheless provides a domain-based classification of protein folds that is straight-forward to map onto known 3D structures in the Protein Data Bank (PDB) [6].

Since it is well known that protein folds are often more evolutionarily conserved than their sequences [7], and since it has been shown that proteins with similar sequences often interact in similar ways [8], it is natural to suppose that close structural homologues should also interact in similar ways. Indeed, several studies have found that the locations of protein interaction sites are often conserved, especially within domain families, regardless of the structures of their binding partners [9, 10, 11, 12]. Additionally, it has also been observed that many protein families employ

only one or a small number of binding sites [13, 14], suggesting that the same surface patch is often re-used. Furthermore, it has been demonstrated previously that the structure of an unknown protein complex may often be successfully modeled using the known binding sites of homologous domains [15, 16]. This may be described as template-based docking, or docking by homology [11, 17].

In order to exploit the above observations, we developed Kbdock to compare and cluster the 3D structures of known DDIs and to provide a systematic way to find structural templates for docking by homology [18, 19]. Essentially, Kbdock is a dedicated relational database which combines the Pfam domain classification with coordinate data from the PDB to analyse and model domain-domain interactions (DDIs) in 3D space.

The Kbdock database can be queried using Pfam domain identifiers, protein sequences, or 3D protein structures. For a given query domain or pair of domains, Kbdock retrieves and displays a non-redundant list of homologous DDIs or domain-peptide interactions (DPIs) in a common coordinate frame. Kbdock may also be used to search for and visualise interactions involving different but structurally similar Pfam families. Thus, structural DDI templates may be proposed even when there is little or no sequence similarity to the query domains.

A fundamental concept in Kbdock is the notion of a protein domain family binding site (DFBS). If one extracts all of the structures from the PDB that involve a given Pfam domain, and if one superposes all such structures onto a representative example of the chosen domain, it is often found that the interaction partner domains of the domain of interest are distributed around just one or a small number of binding sites on the given domain. If the various interaction partners are clustered in 3D space, each cluster may then be used to describe a common family-level binding site on the domain of interest (see Note 4.1). In other words, a DFBS is an abstract representation of all 3D binding-site instances located at the same position within a given domain family. As a natural extension of this idea, we then define a domain family interaction (DFI) as an interaction between two DFBSs. Thus a DFI is the abstract representation of all DDI instances that involve the same pair of DFBSs on the two interacting domain families [18]. This gives a way to define and compare DDIs at a structural level, without needing to be concerned with the precise nature of the residue-residue contacts that might occur within a particular interface between two domains [20]. Indeed, the notion and use of DFBSs and DFIs provides a clear separation between Kbdock and other structural DDI databases such as 3DID [21] and Interactome3D [22].

2 Materials

2.1 The Kbdock Database

The Kbdock database has been described previously [18, 23]. Briefly, Kbdock combines information extracted from the Pfam protein domain classification [24] with coordinate data for structural DDIs from the Protein Data Bank (PDB) [6]. Each DDI is classified as “intra” or “inter” and “homo” or “hetero” according to whether the interaction is within one chain or across two chains, and whether it involves the same or different chains, respectively. The current version of Kbdock uses Pfam version 27.0 and a snapshot of the PDB that was taken in June 2013. After duplicate or near-duplicate interactions are removed (see Note 4.2), the Kbdock database contains a total of 4,001 Pfam DFBSs located on 2,153 different Pfam domains or families, and involved in a total of 5,139 non redundant DDIs. As two or more non redundant DDIs can still correspond to the binding between the same two Pfam domains at the same binding sites, the 5139 non redundant hetero DDIs have been mapped to a total of 3084 distinct DFIs. A full dump of the database is available from the Kbdock web site (*http://kbdock.loria.fr/download.php*).

2.2 The Kbdock Web Interface

Kbdock is normally used via its on-line interface (*http://kbdock.loria.fr/*) [18, 23], although it may also be queried programmatically by expert users in order to execute complex or specialised queries. Here, we describe only the features of Kbdock that are publicly and freely available to the community via the Kbdock web site. This web site has been tested using a range of popular web browsers such as Firefox, Safari, Chrome, and Explorer. Most queries are executed in just a few seconds or less. Thus, there are no login requirements, and all results are presented to the user as new web pages which are generated on the fly. Most results pages link out to the Pfam web site (*http://pfam.xfam.org/*) to allow the user to see detailed descriptions and references for the domains of interest. DDIs stored in Kbdock may be visualised as a network and navigated interactively using the Cytoscape plugin [25].

2.3 3D Visualisation

To support on-line 3D visualisation of results, Kbdock currently uses the Java-based "Jmol" web plugin (*http://jmol.sourceforge.net/*) and, optionally, the more recent Javascript-based "JSmol" plugin (*http://jsmol.sourceforge.net/*). These may easily be installed from the user's web browser. Additionally, Kbdock allows the results of all queries involving 3D structures to be downloaded to the user's workstation and visualised using a high quality 3D molecular visualisation program such as "VMD" [26] or "PyMOL" [27]. Command scripts for these programs may be downloaded (*http://kbdock.loria.fr/download.php*) which let the user view the retrieved structures in high resolution with a minimum of effort.

3 Methods

This section describes various ways in which the Kbdock web site may be browsed and queried. Additional help and examples are available on-line at *http://kbdock.loria.fr/help.php*.

3.1 Browsing the Kbdock Database

Probably the easiest way to learn and understand the Kbdock web interface is to browse the database. If the user selects the *Browse* button at the top of the main Kbdock web page, he is then presented with a short form to choose which category of interaction to browse: inter-chain hetero DDIs; inter-chain homo DDIs; intra-chain hetero DDIs; intra-chain homo DDIs. The user may also choose to browse inter-chain or intra-chain DPIs. The default choice is to browse inter-chain hetero DDIs. Pressing the *Show Pfam families* button then leads to a new page which tabulates the contents of the database for the chosen category. This table shows the total number of DDIs for each Pfam family, the number of representative DDIs (see Note 4.3), and the number of DFBSs within each family.

For example, the row beginning with *Asp* indicates that this domain family (Pfam accession code PF00026) has a total of 19 DDIs which together may be described by six representative DDIs and five DFBSs. Clicking on the *Pfam AC* link for this domain (PF00026) links out to the Pfam entry for *Asp* (*http://pfam.xfam.org/family/PF00026*). This Pfam entry reports that domains in the *Asp* (aspartate protease) family generally have two highly conserved catalytic aspartate residues in an active site cleft that lies between two lobes which appear to be related to each other by a gene duplication event, and that these enzymes typically cleave a peptide substrate which sites in the active site cleft. On the other hand, clicking on the *show* link in the *DDI (REP)* column for *Asp* leads to a new Kbdock page which allows the user to view the six representative DDIs graphically. This page shows the PDB structure codes, chain identifiers, start and end residue numbers, and the Pfam IDs of the six representative DDIs (Figure 1). This page also shows a

Jmol window containing those DDIs superposed using the coordinates of the individual *Asp* domains. It can be seen that Kbdock contains DDIs involving *Asp* and three different protease inhibitor families, namely *Inhibitor_I34*, *Pepsin-I3* and *Serpin*. It can also be seen that *Asp* also has interactions for which structures exist with the *SH3_1* and the antibody *V-set* domains (with the *V-set* interactions being mediated by two distinct DFBSs).

[Figure 1 **near here]**

It is also possible to browse DDIs using the Cytoscape plugin. For example, if the user selects the *Network* button at the top of the main Kbdock web page, he is then presented with a short form to specify the principal Pfam domain of interest and to choose which category of DDI to browse. By default, the Cytoscape plugin shows interaction networks to a depth of two interactors with respect to a given domain. Figure 2 shows a screen-shot of the DDI network that is presented when the user chooses to view the inter-chain hetero interactions that involve the *Asp* domain (PF00026). This network shows the five different domains which interact directly with *Asp*, namely *Inhibitor_I34*, *Pepsin-I3*, *Serpin*, *SH3_1*, and *V-set*, along with all of the DDI partners of those five (the majority of which involve interactions with the large antibody *V-set* family).

[Figure 2 **near here]**

3.2 Domain-Peptide Interactions

Kbdock's network view provides a convenient and rapid way to see which domains in a protein interaction network have 3D structures. However, because DDIs and DPIs are treated separately in both Kbdock and Pfam, it is often advisable to perform a separate search for DPIs for the domain of interest. For example, searching for DPIs with the *Asp* domain as query retrieves two representative interactions involving the pro-enzyme forms of two aspartate proteases, in which the active site is blocked by the short *A1_propeptide* motif, as shown in Figure 3. It should be noted that this figure shows two different DFBSs on the protease. The first DFBS, extracted from PDB structure 1HTR, shows the propeptide blocking the binding site cleft of the protease. This binding mode may be considered as the "true" biological interaction. The second DFBS, extracted from PDB structure 3VCM, shows a smaller contact somewhat away from the protease active site cleft. This secondary contact is most probably a non-biological crystal contact which arises from the fact that the *Asp* domains often crystallise as homodimer structures. Note 4.4 provides some additional remarks on distinguishing biological from non-biological contacts.

[Figure 3 **near here]**

3.3 Structural Neighbour Interactions

It can sometimes be interesting to view structural neighbour interactions of a given domain, either because relatively few DDIs exist for the domain of interest, or because one wishes to explore possible structural homologies which might not be detected by conventional sequence alignment searches. For each Pfam domain for which structural interactions exist, Kbdock maintains a list of similar structures from different Pfam domains which have been found by our "Kpax" structural alignment program [28] (see Note 4.5). Then, using these lists, Kbdock searches for and retrieves structural neighbour interactions in the same way as for DDIs that directly involve the given query structure(s). For example, the results page mentioned above for *Asp* DDIs shows that two inter-chain hetero and two intra-chain homo DDIs exist for structural neighbours of the *Asp* query domain, both involving the *TAXi_N* and *TAXi_C* xylanase inhibitor domains. There also exist three inter-chain homo and one intra-chain homo DDIs, all of which involve the *RVP* (retroviral aspartyle protease) domain (PF00077).

Following the link for the representative intra-chain homo interaction with *RVP* shows that the representative structure (PDB code 4EP3) for this domain superposes very well onto the N-

terminal lobe of the representative structure for *Asp* (PDB code 4D8C) with 13 sequence identities out of 83 aligned residues (15.7% identity) and with an aligned root mean squared deviation (RMSD) of 2.29 Å. This superposition supports the proposition that the *Asp* and *RVP* families are evolutionarily related, as described in more detail on the Pfam web site (*http://pfam.xfam.org/family/PF00077*).

On the other hand, following the link for the representative inter-chain hetero interactions, it can be seen (Figure 4) that the *TAXi_N* and *TAXi_C* domains superpose very well onto the N-terminal and C-terminal lobes of *Asp*, respectively. Indeed, the superposition of *TAXi_N* from PDB structure 3AUP onto the representative *Asp* structure (4D8C) gives 112 aligned residues with 21 sequence identities (18.7% identity), with an aligned RMSD of 2.76Å. The corresponding superposition of *TAXi_C* onto *Asp* using the same PDB structure gives 19 identities out of 129 aligned residues (14.7%) with a of 2.23Å. These very tight superpositions strongly suggest that these xylanase inhibitor domains are also evolutionarily related to the *Asp* family.

[Figure 4 **near here]**

3.4 Searching for DDI Docking Templates

Because one of the principal aims of Kbdock is to be able to exploit existing 3D structures to find candidate templates with which to model an unsolved complex, Kbdock naturally supports queries involving a pair of sequences or structures which are presumed to interact, or "dock". To support searching for docking templates, the user may query Kbdock by pasting two amino acid sequences into a query form or by uploading two 3D protein structures. In either case, Kbdock uses the "PfamScan" utility [24] to identify the Pfam domains within the given sequences or structures, and it then asks the user to select which structures should be considered as queries for the docking template search.

As a worked example, we will consider the arrowhead protease inhibitor A (API-A) enzyme-inhibitor complex, which was presented to the docking community as target 40 in Round 18 of the CAPRI blind docking experiment [29]. This target is a complex between API-A and two trypsin molecules [30]. At the time that this target was first presented to the CAPRI predictors, the Kbdock database had not yet been implemented. Nonetheless, it is an interesting complex to consider because it allows the capabilities of Kbdock to be demonstrated easily.

If the user navigates to the *Search* page on the Kbdock web site, and then selects the option *Identify Pfam domains for a given structure*, he can upload the 3D structure files for target 40 that were provided by the CAPRI organisers (comprising the API-A protease inhibitor and two trypsins). Selecting *Continue* then takes the user to a results page which shows that his PDB files contain three domains, namely *Kunitz_legume* (PF00197) and two copies of *Trypsin* (PF00089), which were found automatically using the PfamScan utility. In this page, the Pfam AC numbers are presented as active links to the corresponding pages on the Pfam web site. These links allow the user to view more detailed information and references about the query domains in a fresh browser window or tab.

Returning to the results page, if the user checks the selection button next to *Kunitz_legume* and one of the two *Trypsin*s, he may then press the *Find Templates* button to search for existing DDIs which could serve as a 3D docking template for the selected pair of domains. Kbdock then presents a summary page which shows that a total of eight DDIs involving *Kunitz_legume* and *Trypsin* are available, and that these interactions may be described by two representative DDIs. Clicking on the *show all* link then leads to a results page (Figure 5) which shows the selected interactions superposed in a Jmol window. In this figure, it can be seen that a trypsin from PDB structure 3E8L occupies one binding site on the *Kunitz_legume* domain (arbitrarily numbered DFBS 1 by Kbdock), while the remaining seven trypsins (extracted from other non-redundant instances of PDB structures) occupy another *Kunitz_legume* binding site (DFBS 2). In other

words, it may be observed that the majority of the *Kunitz_legume* inhibitors use the same surface loop region to bind to trypsin but at least one member of this family binds trypsin *via* a different surface loop.
In fact, the PDB structure 3E8L is the published solution structure for CAPRI target T40 [30]. Thus, at the time that this target was presented to the CAPRI predictors, no structural template was available for the DFBS 1 interaction. Nonetheless, we correctly predicted the second API-A inhibitory loop based on its structural similarity to the known binding site loop (DBFS 2) [31]. This demonstrates that retrieving and analysing the structures of existing DDIs can provide useful clues or hypotheses for the prediction of new interactions.
Of course, because today both of the above DBFSs exist in Kbdock, we now have a richer set of templates with which to model other new interactions involving the same domain families. Furthermore, even in cases where DDI templates do not exist for precisely the same Pfam families of a docking target, we showed recently that structural neighbour DDIs can provide a useful additional source of docking templates [23]. We therefore encourage the user to consider this possibility when using Kbdock to model protein complexes by homology.
[Figure 5 **near here]**

4 Notes

4.1 How DFBSs Are Defined

The Kbdock database is populated using a number of in-house scripts [18, 23]. For every protein chain in the PDB, its sequence is processed by PfamScan in order to cut the chain into separate domains. Then, using the same criteria as Stein *et al.* [21], each domain having five or more atomic contacts (i.e. van der Waals contacts, hydrogen bonds, or salt-bridges) with another domain is considered to participate in a DDI, and each DDI is classified as “intra” or “inter” and “homo” or “hetero” according to whether the interaction is within one chain or across two chains, and whether it involves the same or different chains, respectively. Each domain is annotated with secondary structural information from the “DSSP” program [32] For each Pfam family, all of the domains of a given interaction type are then aligned and superposed along with their interaction partners using our Kpax structural alignment program in order to place all related DDIs into a common coordinate frame. For each such DDI, a vector is calculated between the centre of the domain of interest and a weighted average of its interface residues. These vectors are then clustered in order to define shared binding sites on the domain, irrespective of the type of binding partner. We call each such distinct cluster a DFBS, as it represents a binding site that is common to all domains within the given Pfam family regardless of the nature of the residues in any particular instance of a DDI.
Within the Kbdock database, each DFBS is identified by its Pfam family identifier and a numerical identifier arising from the clustering step. Thus, each DFBS is essentially a composite database key, and each DDI involves a pair of such keys. Consequently, DDIs may be retrieved and manipulated very efficiently, which led us to propose a systematic case-based reasoning approach for docking by homology [19].

4.2 Filtering Duplicate Structures and Interactions

Many of the DDIs extracted directly from PDB structures are redundant, either because a single crystal structure may contain several symmetry-mates, or because a given complex may have been solved several times under different crystallographic conditions, for example. Therefore, to achieve a robust classification and reliable statistics, Kbdock eliminates redundant DDIs by applying the NRDB90 program [33] with a threshold of 99% sequence identity to the entire set of

sequences built from the concatenation of the two interacting domain sequences in each DDI. This filtered set of DDIs is then clustered using our binding site direction vector algorithm in order to define the DFBSs. Finally, the DDI instances involving each DFBS are filtered again using a 60% sequence similarity threshold in order to retain mostly distinct pairs of domains associated with any given DFBS.

4.3 Representative Structures

Because some Pfam domains can have many 3D structures in the PDB that have interactions with other domains, it can be difficult and slow to visualise all of the relevant structures together, even after obvious duplicate structures have been removed (Section 4.2). Therefore, when Kbdock initially clusters DDIs to define the binding sites within each Pfam family, it selects a single representative example for each of the four interaction types (hetero/homo–inter/intra). More specifically, since each DFBS is defined as a cluster of binding site vectors, Kbdock selects the domain instance whose binding site vector lies closest to the average of all vectors as the representative 3D structure for that domain family.

4.4 Distinguishing Biological and Crystallographic Contacts

When browsing structural databases such as Kbdock, or indeed the PDB itself, it is easy to forget that many 3D protein structures derive from regular crystal structures which can have multiple domain-domain contacts, and that it is often difficult to discern which, if any, contacts correspond to *in vivo* biological interactions, and which contacts are merely artefacts of the crystal packing. Furthermore, even if it might be known that a given protein exists *in vivo* as a homodimer, for example, this knowledge is often not apparent from the annotations or coordinates in a PDB file [34]. Consequently, Kbdock does not attempt to distinguish "true" biological interfaces from crystal contacts, and it therefore collects and reports all observed contacts according to the criteria described above. It has been noted previously that interfaces with large surface areas often correspond to the true biological interfaces, but this rule of thumb does not hold in every case [34]. Thus, if Kbdock reports two or more interactions involving the same pair of domains, the user is advised to download and examine the original PDB files and references in order to try to distinguish "true" biological interactions from crystallographic artefacts.

4.5 Structural Neighbour Interactions

Kbdock uses our Kpax structural alignment program to calculate a list of structural neighbours for the members of each Pfam family. This list is then cross-checked with Kbdock's table of DDIs in order to provide a pre-calculated list of "structural neighbour" interactions – i.e. DDIs which are structurally similar to the query domains, but which do not belong to exactly the same Pfam domain as the query. Kpax measures structural similarity using a normalised Gaussian overlap score calculated between aligned pairs of atom coordinates. In Kbdock, any pair of domains that give a Kpax similarity score of 0.25 or greater are assumed to be structurally similar (i.e. they have largely the same overall fold). The Kpax program may be downloaded for academic use at *http://kpax.loria.fr/*.

Acknowledgments

This work was funded in part by the Agence Nationale de la Recherche, grant reference numbers ANR-08-CEXC-017-01 and ANR-MNU-006-02.

References

[1] Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML (2009). Progress and challenges in predicting protein-protein interaction sites. *Briefings in Bioinformatics*, 10(3):233–246

[2] Berman HM (2008). The protein data bank: a historical perspective. *Acta Crystallographica*, A38:88–95

[3] Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A (2010). The Pfam protein families database. *Nucleic Acids Research*, 38:D211–D222

[4] Murzin AG, Brenner SE, Hubbard T, Chothia C ()1995). SCOP – a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540

[5] Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, and Orengo CA (2009). The CATH classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37:D310–D314.

[6] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C (2002). The protein data bank. *Acta Crystallographica Section D-Biological Crystallography*, 58:899–907.

[7] Chothia C, Lesk AM (1986). The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5:823–826.

[8] Aloy P, Ceulemans H, Stark A, Russell RB (2003). The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*, 332(5):989–998

[9] Keskin O, Ma BY, Nussinov R (2005). Hot regions in protein-protein interactions: The organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, 345(5):1281–1294

[10] Korkin D, Davis FP, Sali A (2005). Localization of protein-binding sites within families of proteins. *Protein Science*, 14:2350–2360

[11] Korkin D, Davis FP, Alber F, Luong T, Shen MY, Lucic V, Kennedy MB, Sali A(2006). Structural modeling of protein interactions by analogy: application to PSD-95. *PLoS Computational Biology*, 2(11):e153

[12] Gunther S, May P, Hoppe A, Frommel C, Preissner R (2007). Docking without docking: ISEARCH – prediction of interactions using known interfaces. *Proteins: Structure Function and Bioinformatics*, 69(4):839–844

[13] Shoemaker BA, Panchenko AR, Bryant SH (2006). Finding biologically relevant protein domain interactions: Conserved binding mode analysis. *Protein Science*, 15(2):352–361

[14] Keskin O, Nussinov R (2007). Similar binding sites and different partners: Implications to shared proteins in cellular pathways. *Structure*, 15:341–354.

[15] Launay G, Simonson T (2008). Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC Bioinformatics*, 9:427

[16] Kundrotas PJ, Lensink MF, Alexov E (2008). Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles. *International Journal of Biological Macromolecules*, 43(2):198–208

[17] Kundrotas PJ, Alexov E (2006). Predicting 3D structures of transient protein-protein complexes by homology. *BBA - Proteins & Proteomics*, 1764(9):1498–1511

[18] Ghoorah AW, Devignes M-D, Smail-Tabbone M, Ritchie DW (2011). Spatial clustering of protein binding sites for template based protein docking. *Bioinformatics*, 27(20):2820–2827

[19] Ghoorah AW, Smail-Tabbone M, Devignes M-D, Ritchie DW (2013). Protein docking using case-based reasoning. *Proteins: Structure Function And Genetics*, 81:2150–2158

[20] Ghoorah AW, Devignes M-D, Alborzi S-Z, Smail-Tabbone M, Ritchie DW (2015). A structure-based classification and analysis of protein domain family binding sites and their interactions. *Biology*, 4:327–343

[21] Stein A, Ceol A, Aloy P (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 39:D718–D723

[22] Mosca R, Ceol A, Aloy P (2013). Interactome3D: adding structural details to protein networks. *Nature Methods*, 10(1):47–53

[23] Ghoorah AW, Devignes M-D, Smail-Tabbone M, Ritchie DW (2014). KBDOCK 2013: a spatial classification of 3D protein domain family interactions. *Nucleic Acids Research*, D42:389–395

[24] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy DR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014). Pfam: The protein families database. *Nucleic Acids Research*, D42:220–230

[25] Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T (2012). A travel guide to Cytoscape plugins. *Nature Methods*, 9(11):1069–1076

[26] Humphrey W, Dalke A, Schulten K (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38

[27] Schrödinger LLC (2010). The PyMOL molecular graphics system, version 1.3r1. http://www.schroedinger.com. Accessed 10 July 2015
[28] Ritchie DW, Ghoorah AW, Mavridis L, Venkatraman V (2012). Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, 28:3274–3281.
[29] Janin J (2010). The targets of CAPRI rounds 13–19. *Proteins: Structure, Function, Bioinformatics*, 78:3067–3072
[30] Bao R, Zhou C-J, Jiang C, Lin S-X, Chi C-W, Chen Y (2009). The ternary structure of the double-headed arrowhead protease inhibitor API-A complexed with two trypsins reveals a novel reactive site conformation. *Journal of Biological Chemistry*, 284:26676–26684
[31] Lensink MF, Wodak SJ (2010). Docking and scoring protein interactions: Capri 2009. *Proteins: Structure Function and Bioinformatics*, 78(15):3073–3084
[32] Kabsch W, Sander C (1983). Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637
[33] Holm L, Sander C (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, 14(5):423–429
[34] Krissinel E, Henrick K (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372(3):774–797

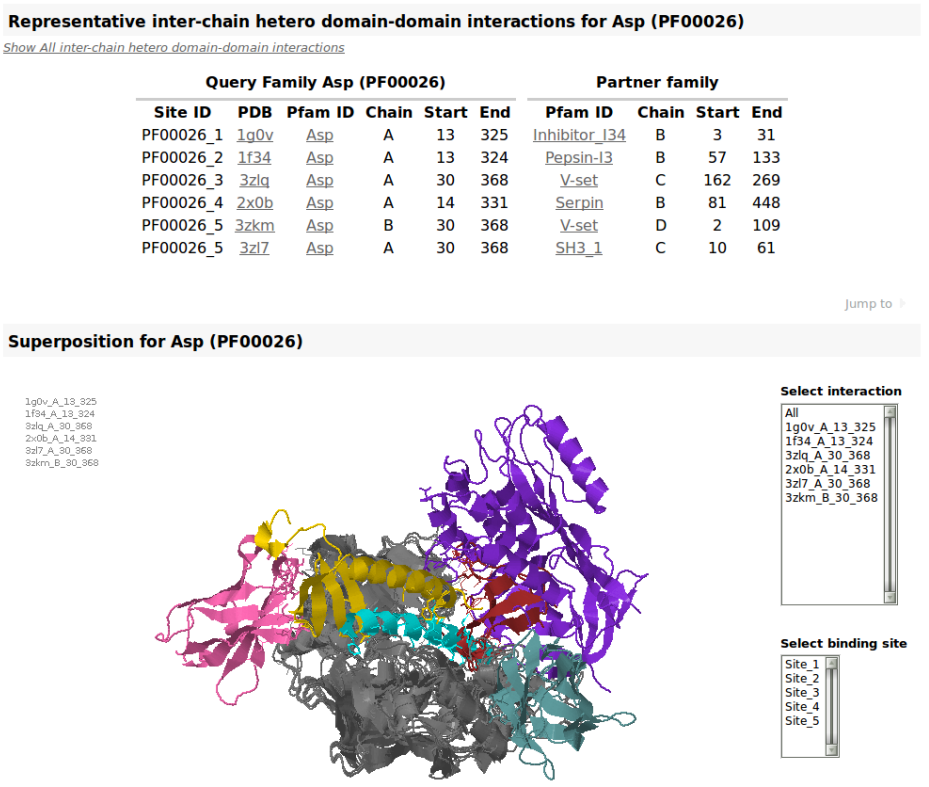

Figure 1: Screen-shot of part of the Kbdock results page that is displayed for the six representative interactions involving the *Asp* (PF00026) domain. The six *Asp* domains are superposed in grey. *Inhibitor_I34* is shown in cyan, *Pepsin-I3* in yellow, *Serpin* in purple, *SH3_1* in red, and two different antibody *V-set* domain interactions are shown in pink and blue. The Kbdock results page also contains an annotated multiple sequence alignment of the *Asp* domains, which is not shown here.

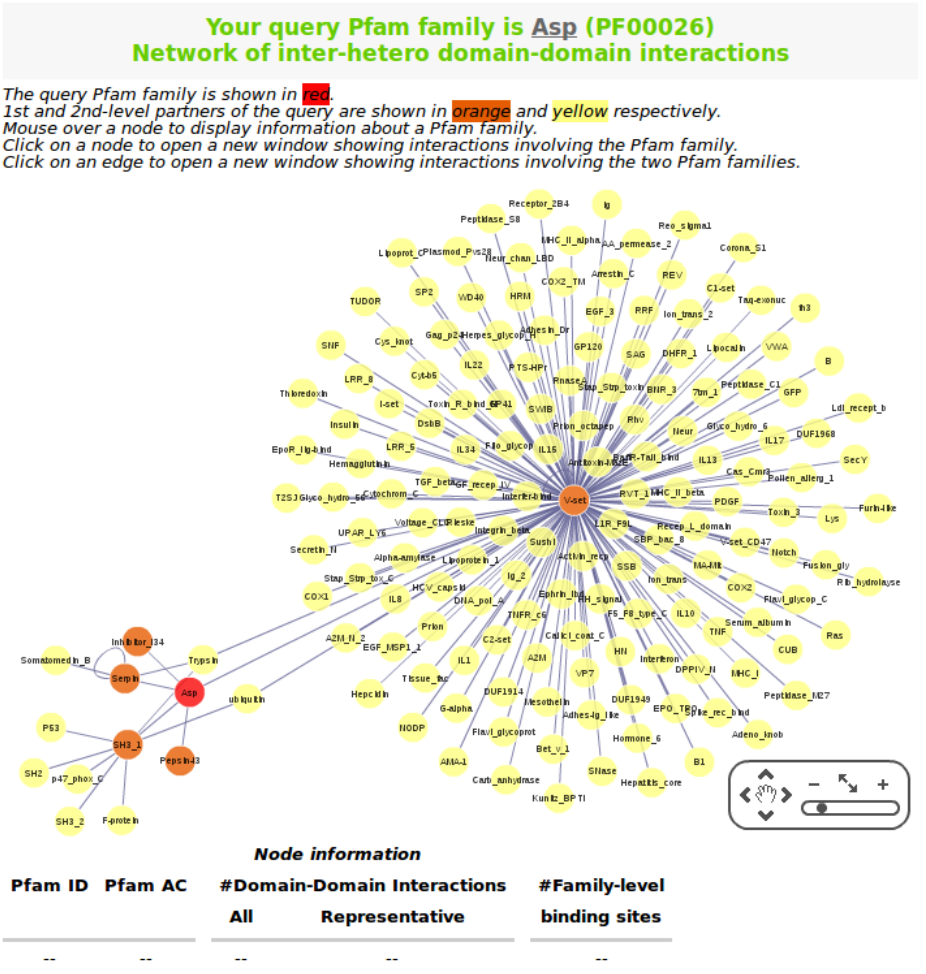

Figure 2: Screen-shot of the DDI network involving the *Asp* (PF00026) domain, drawn using the Cytoscape plugin. Here, the *Asp* domain is shown as a red circle. The five domains that interact with *Asp* are shown in orange (*Inhibitor_I34*, *Pepsin-I3*, *Serpin*, SH3_1, and *V-set*), and all domains having additional interactions with those five interactors are shown as yellow circles. Moving the mouse cursor over a domain will cause some details about that domain to replace the dashes at the bottom of the image. Clicking on a domain will cause a new Kbdock window to appear in which the selected domain is treated as a new query for which its interaction partners are shown. Similarly, clicking on an edge between two domains will generate a new Jmol window which shows the interaction in 3D.

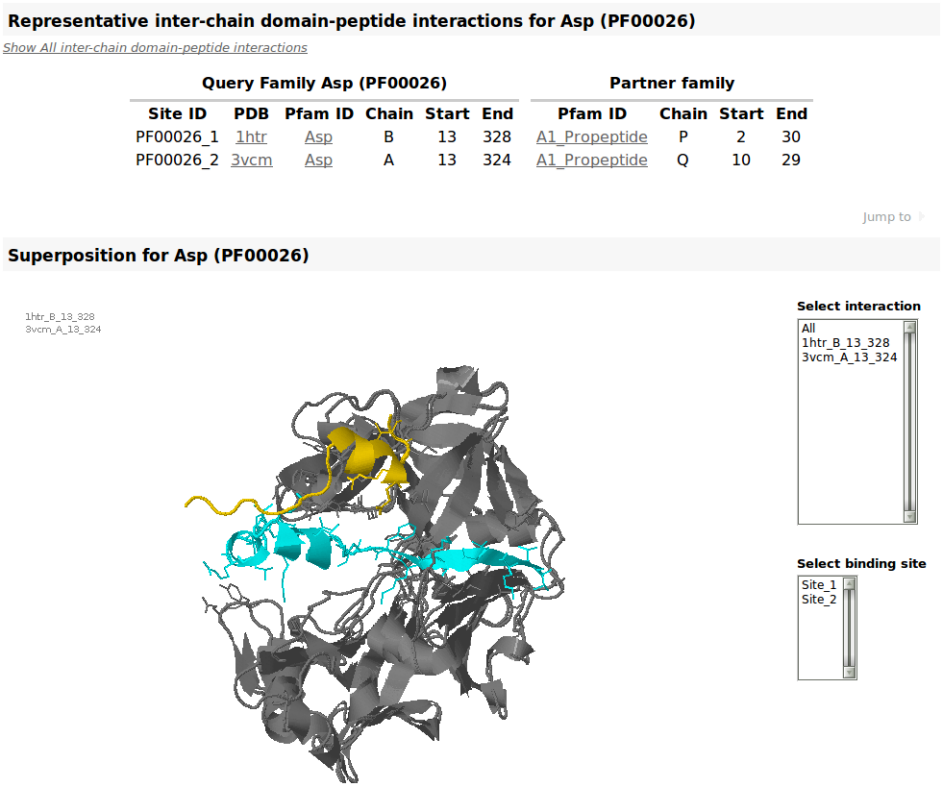

Figure 3: Screen-shot of part of the Kbdock results page that is displayed to show the two DPIs involving the *Asp* (PF00026) domain (shown in grey). The first DPI (PDB code 1HTR) is shown in cyan, and the second DPI (PDB code 3VCM) is shown in yellow. Because the coordinates provided in the two PDB files show that both PDB structures were solved as homo-dimers, and because the interface in 1HTR is much more extensive than in 3VCM, it may be supposed that the former interface corresponds to the "true" biological interface, whereas the latter represents a non-biological crystallographic contact. Note that the peptide colours in this image are not related to those of Figure 2. 2

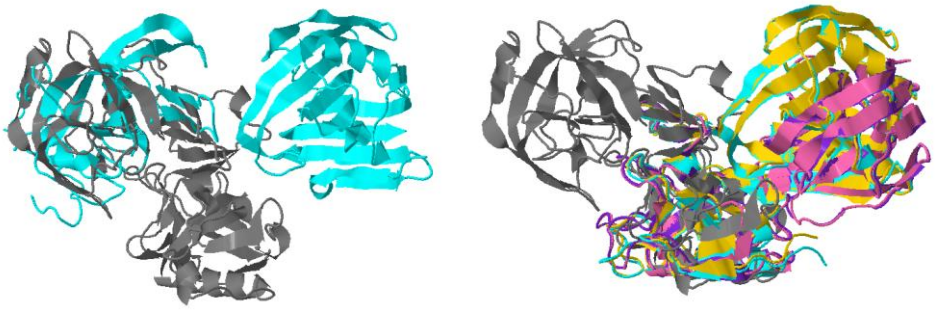

Figure 4: Kbdock superpositions of the *Asp* domain (PF00026) with its nearest structural neighbour domains, *TAXi_N* (PF14543) and *TAXi_C* (PF14541), found by Kpax. The image on the left shows the superposition of the *TAXi_N* domain onto the N-terminal domain of *Asp* drawn in grey using PDB structure 4D8C as the representative structure for *Asp*, along with its DDI partner domain *Glyco_hydro_11* (PF00457) drawn in cyan using PDB structure 1T6G. The image on the right shows the superposition of four *TAXi_C* domains onto the C-terminal domain of *Asp* drawn in grey using PDB structure 4D8C, along with its DDI partner domains *Glyco_hydro_11* (cyan: PDB code 2B42; gold: PDB code 3HD8) and *Glyco_hydro_12* (PF01670; pink: PDSB code 3VLB, chain A; violet: PDB code 3VLB, chain C). These tight superpositions strongly suggest that the *TAXi_N* and *TAXi_C* domains are evolutionarily related to *Asp*.

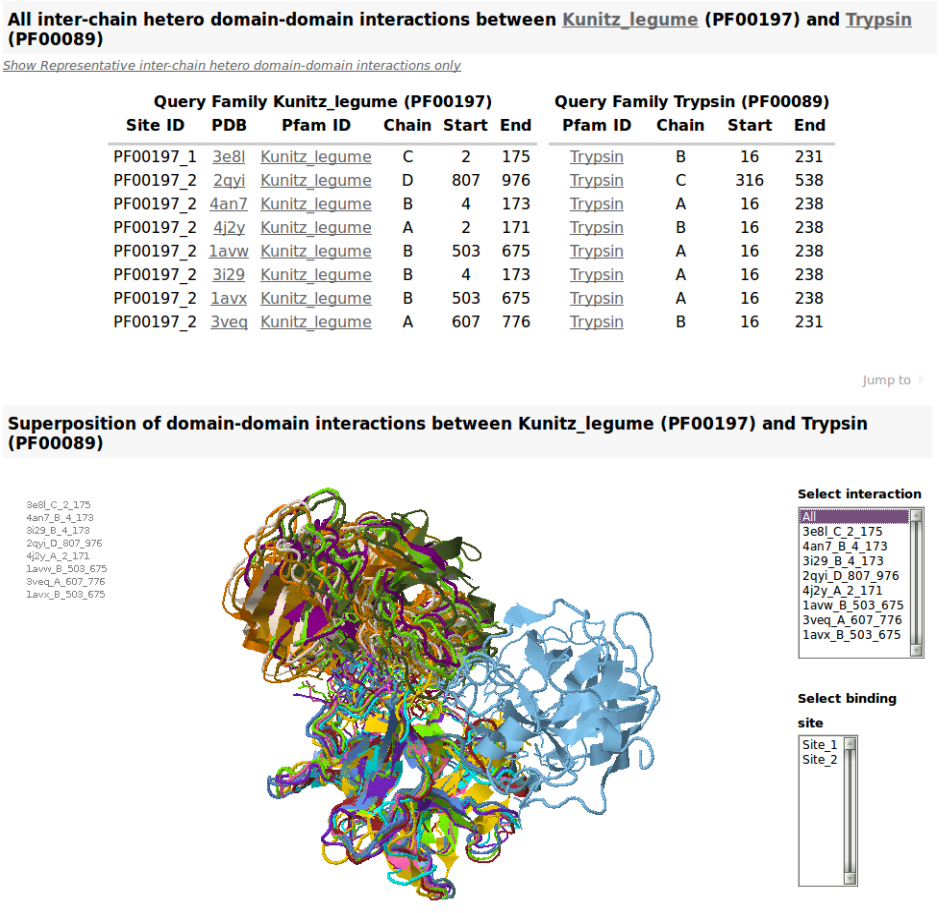

Figure 5: Screen-shot of the Kbdock results page shown after searching for interactions involving the *Kunitz_legume* (PF00197) and *Trypsin* (PF00089) domains. In this figure, eight *Kunitz_legume* domains are superposed to reveal that seven of the *Trypsin* domains occupy the same binding site (DFBS 2 in Kbdock), while in the 3E8L structure another trypsin occupies a different binding site (DFBS 1). In fact, the PDB structure 3E8L contains the solution structure for CAPRI target 40, namely the API-A/trypsin complex in which one API-A protein binds two trypsins simultaneously using the two DFBSs shown here. Therefore, at the time that this target was presented to the CAPRI predictors, a structural template was available for the DFBS 2 interaction, but not for DFBS 1.